COLLA 2022

The Twelfth International Conference on Advanced Collaborative Networks, Systems and Applications

ISBN: 978-1-61208-976-8

May 22nd –26th, 2022

Venice, Italy

COLLA 2022 Editors

Samia Ben Rajeb, Université Libre de Bruxelles, Belgium

COLLA 2022

Forward

The Twelfth International Conference on Advanced Collaborative Networks, Systems and Applications (COLLA 2022) continued a series of events dedicated to advanced collaborative networks, systems and applications, focusing on new mechanisms, infrastructures, services, tools and benchmarks.

Collaborative systems have raised to become an inherent part of our lives, supported by global infrastructures, technological advancements and growing needs for coordination and cooperation. While organizations and individuals relied on collaboration for decades, the advent of new technologies (e.g. from wikis to real-time collaboration, groupware to social computing, service-oriented architecture to distributed collaboration) for inter- and intra- organization collaboration enabled an environment for advanced collaboration. As a consequence, new developments are expected from current networking and interacting technologies (protocols, interfaces, services, tools) to support the design and deployment of scalable collaborative environments. Current trends include innovations in distributed collaboration, collaborative robots, autonomous systems, online communities, or real-time collaboration protocols.

We take here the opportunity to warmly thank all the members of the COLLA 2022 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to COLLA 2022. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the COLLA 2022 organizing committee for their help in handling the logistics of this event.

COLLA 2022 Chairs

COLLA 2022 Steering Committee

Mudasser F. Wyne, National University, USA
Yiannis Koumpouros, University of West Attica, Greece
Bruno Vallespir, Université de Bordeaux, France

COLLA 2022 Publicity Chairs

Mar Parra, Universitat Politècnica de València (UPV), Spain
Hannah Russell, Universitat Politècnica de València (UPV), Spain

COLLA 2022
Committee

COLLA 2022 Steering Committee

Mudasser F. Wyne, National University, USA
Yiannis Koumpouros, University of West Attica, Greece
Bruno Vallespir, Université de Bordeaux, France

COLLA 2022 Publicity Chairs

Mar Parra, Universitat Politècnica de València (UPV), Spain
Hannah Russell, Universitat Politècnica de València (UPV), Spain

COLLA 2022 Technical Program Committee

Jameela Al-Jaroodi, Robert Morris University, USA
Siva Ariram, University of Oulu, Finland
Benjamin Aziz, University of Portsmouth, UK
Samia Ben Rajeb, University of Brussels, Belgium
Lasse Berntzen, University of South-Eastern Norway, Norway
Ramon Chaves, Federal University of Rio de Janeiro, Brazil
Richard Chbeir, Université de Pau et des Pays de l'Adour (UPPA), France
Feiching Chen, National Central University, Taiwan
Ioannis Chrysakis, FORTH-ICS, Greece / Ghent University, IDLab, imec, Belgium
António Correia, University of Trás-os-Montes e Alto Douro | INESC TEC, Portugal
Ireneusz Czarnowski, Gdynia Maritime University, Poland
Antonio De Nicola, ENEA, Italy
Arianna D'Ulizia, National Research Council - IRPPS, Italy
Christian Esposito, University of Napoli Federico II, Italy
Michael Fuchs, Wilhelm Büchner University of Applied Sciences, Germany
Dimitrios Georgakopoulos, Swinburne University of Technology, Australia
Juscimara Gomes Avelino, Federal University of Pernambuco, Brazil
António Guilherme Correia, INESC TEC, Portugal
Atsuo Hazeyama, Tokyo Gakugei University, Japan
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Takeshi Ikenaga, Kyushu Institute of Technology, Japan
Ilias Karasavvidis, School of the Humanities Argonafton & Filellinon, Greece
Hassan A. Karimi, University of Pittsburgh, USA
Basel Katt, Norwegian University of Science and Technology, Norway
Yiannis Koumpouros, University of West Attica, Greece
Thomas Lampoltshammer, Danube University Krems, Austria
Madina Mansurova, Al-Farabi Kazakh National University, Kazakhstan
José Martins, INESC TEC & Polytechnic of Leiria, Portugal
Dhruv Mehta, Aument, Austria
Michele Melchiori, Università degli Studi di Brescia, Italy
Fernando Moreira, Universidade Portucalense, Portugal

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Automated Data Preprocessing for Machine Learning Based Analyses

Akshay Paranjape
IconPro GmbH
Aachen, Germany
akshay.paranjape@iconpro.com

Praneeth Katta
IconPro GmbH
Aachen, Germany
praneeth.katta@iconpro.com

Markus Ohlenforst
IconPro GmbH
Aachen, Germany
markus.ohlenforst@iconpro.com

***Abstract*—Data preprocessing is crucial for Machine Learning (ML) analysis, as the quality of data can highly influence the model performance. In recent years, we have witnessed numerous literature works for performance enhancement, such as AutoML libraries for tabular datasets, however, the field of data preprocessing has not seen major advancement. AutoML libraries and baseline models like RandomForest are known for their easy-to-use implementation with data-cleaning and categorical encoding as the only required steps. In this paper, we investigate some advanced preprocessing steps such as feature engineering, feature selection, target discretization, and sampling for analyses on tabular datasets. Furthermore, we propose an automated pipeline for these advanced preprocessing steps, which are validated using RandomForest, as well as AutoML libraries. The proposed preprocessing pipeline can also be used for any ML-based algorithms and can be bundled into a Python package. The pipeline also includes a novel sampling method - "Bin-Based sampling" which can be used for general purpose data sampling. The validity of these preprocessing methods has been assessed on OpenML datasets using appropriate metrics such as Kullback-Leibler (KL)-divergence, accuracy-score, and r2-score. Experimental results show significant performance improvement when modeling with baseline models such as RandomForest and marginal improvements when modeling with AutoML libraries.**

***Index Terms*—*AutoML; Preprocessing; Feature Engineering; Feature Generation; Feature selection; Sampling*.**

I. Introduction

Data preprocessing is a crucial step in Machine Learning (ML) as the quality of data can have a significant influence on its performance. Preprocessing is performed to prepare the compatible dataset for analysis as well as to improve the performance of the ML model. Preprocessing steps can be roughly categorized into two types: model compatible preprocessing (Type 1) and quality enhancement preprocessing (Type 2). A common example of model compatible preprocessing step is the encoding of string values to either Label Encoded values or One-Hot Encoded values based on the model requirements. Preprocessing steps like data cleaning and missing value imputation fall into the category of Type 1, while other generic preprocessing steps like standardization and normalization, and cyclic transformation fall into Type 2 category. The focus of this paper is Type 2 preprocessing for supervised learning of tabular datasets.

In recent years, the ML field for tabular datasets has been heavily researched for performance enhancement of ML-based models, especially automated Machine Learning (AutoML) [6]- [8]. However only a few papers [1] [2] have investigated advanced Type 2 preprocessing steps (mainly feature generation). The validation of these preprocessing steps together with AutoML libraries has not been studied yet. We have considered three relevant AutoML libraries, namely AutoGluon [8], AutoSklearn [6], and H2O [7]. The different preprocessing steps supported by these AutoML libraries are summarized in Table I. It can be inferred from Table I that advanced preprocessing steps like feature engineering and feature selection have not been implemented in these AutoML libraries. In this paper, we first investigate some advanced Type 2 preprocessing steps and later propose an automated preprocessing pipeline based on our research. A validation study of the proposed pipeline is conducted on both the baseline model and AutoML libraries.

The automated preprocessing pipeline is designed with the main objective of automatically generating new features from existing input features to improve the performance metric. If the dataset size is large, feature engineering can take a longer computational time. Therefore, required research in the sampling field is evident, for which we propose the Bin-Based sampling method as an alternative to Random Sampling. Before proceeding with feature engineering, unnecessary, irrelevant, and highly insignificant features are removed since these features have neutral or significantly low information gain for the target variable. These three techniques, viz. feature engineering, feature selection, and sampling are the main aspects of this paper. Therefore, in Section II, related work for these three techniques is briefly presented. Further, in Section III, methodology is presented. In Section IV, experiments and the results obtained are tabulated. We conclude the work in Section V.

II. Related Work

A. Feature Engineering

Feature engineering is the process of generating new features with the help of domain knowledge. The construction of novel features for the enhancement of predictive learning is time-intensive and often requires field expertise. With the appropriate addition of features, predictive models can show significant performance improvement. Cognitio by Khurana et al. [1] demonstrated a novel method for automated feature engineering in supervised learning. Cognito performs row-wise transforms over instances for all valid features, each

 ISBN: 978-1-61208-976-8

TABLE I
PREPROCESSING STEPS INCLUDED IN DIFFERENT AUTOML SOLUTIONS

Name	AutoSklearn	AutoKeras	TPOT	AutoGluon	H2O
Balancing	yes	no	no	yes	yes
Categorical encoding	yes	yes	yes	yes	yes
Imputation	yes	yes	no	yes	yes
Standardization/Normalization	yes	yes	no	yes	yes
Others	Densifier, PCA, minority coalescence, select percentile	Data augmentation	Feature selector	Introduce "unknown category"	None

producing a new column or columns. The number of possible transformations is an unbounded space considering various combination of features. These function transforms could be unary, binary, or multiple transforms [1]. As the number of transforms can increase exponentially based on the number of input columns, the pruning step is included by Cognito for feature selection to ensure a manageable size of the dataset.

Katz et al. introduced a framework ExploreKit [2] for automated feature generation. Katz et al. demonstrate new feature findings by using unary operators such as inverse, addition, multiplication, division, etc., as well as higher-order operators. The huge number of features generated in ExploreKit are pruned and validated using a Ranking Model. A two-step approach is proposed by ExploreKit where the generated features in the first step are ranked based on meta-features in the second step.

Galhotra et al. [3] focus on an automated method to utilize domain structured knowledge to perform feature addition. They further developed a tool KAFE (Knowledge Aided Feature Engineering) to attain knowledge about similar analyses from 25 million tables available on the internet. Hoag et al. presented a neural network approach to generate new features from relational databases [4]. They use a set of Recurrent Neural Networks (RNNs) that takes as input a sequence of vectors and outputs a vector (with new generated features). The Data Science Machine [5] developed a deep feature engineering algorithm for relational databases and cannot be generalized to tabular datasets.

B. Feature selection

In contrast to feature engineering where new features are generated from the existing ones, feature selection implies selecting useful features from the available set of input features, i.e., a subset of features. Tereno et al. [10] consider various search strategies for feature selection namely heuristic and probabilistic. They illustrate the importance of removing unnecessary features based on class separability measures. The elimination of non-relevant features or features with negligible importance can significantly reduce computation time and resources [10]. Aliferis et al. [11] introduced an algorithmic framework to learn local casual structure for target structure that is later used to select features. The most popular approach for feature selection includes correlation, Bayesian error rate, information gain, entropy measures, etc., [12]. Elssied et al. [13] demonstrate the use of a one-way Analysis of Variance (ANOVA) F-test for feature selection in the context of email spam classification.

C. Sampling

ML algorithms should be trained on a complete dataset because a higher amount of data can improve performance. But a sample set can help get a quick overview of data quality as well as determine its characteristics. The popular approach of sampling is Random sampling with or without replacement [15]. The literature for sampling dates back to 1980 when Cochran [14] first introduced the concept of stratified sampling. Stratified sampling divides samples into homogeneous subgroups and later the data is randomly sampled from these subgroups. Rojas et al. [15] in their survey concluded that the majority of data scientists use random sampling, stratified sampling, or sampling by hand. Section III-B elaborates on the stratified sampling technique in detail.

III. METHODOLOGY

In this paper, we present an automated preprocessing pipeline that includes advanced preprocessing methods which are generally not available in AutoML libraries. The aim of this research is to develop an automated pipeline that can improve the performance of predictive modeling on tabular datasets. The proposed pipeline can be used with any ML algorithms as well as for AutoML libraries. The novelty aspects of this paper are as mentioned below:

- Hybrid Feature Engineering (HFE) method
- Generalized automated preprocessing pipeline
- Sampling technique

The proposed preprocessing pipeline is validated by analyzing its implementation on OpenML datasets [18]. This section consists of four preprocessing steps each representing an element of the proposed pipeline, namely feature selection, sampling, target discretization, and feature engineering. These steps are described in detail below with their pseudo algorithms.

A. Feature selection

Feature selection implies selecting important features from the list of input features or in other words eliminating less significant features. As mentioned in Section II, a popular approach for feature selection is the correlation coefficient. In this paper, we present a mixture of variance and correlation analysis for feature selection. Inspired by [10], feature selection has been categorized into three parts:

- removal of redundant features

 ISBN: 978-1-61208-976-8

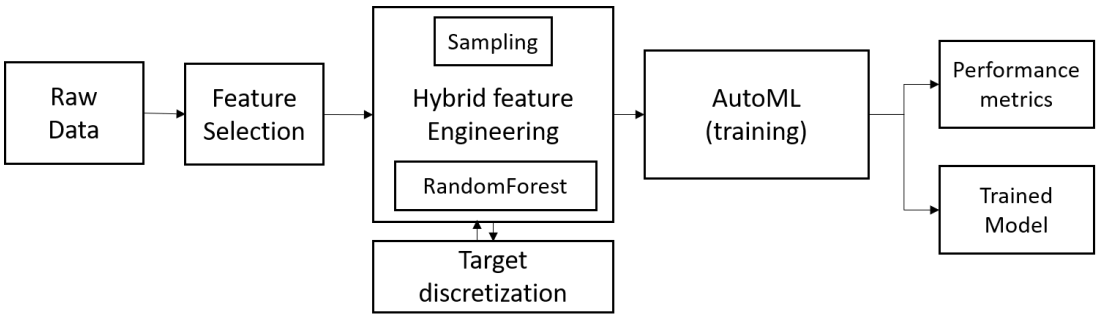

Fig. 1. Block diagram for preprocessing pipeline together with the integration of AutoML module

- removal of highly correlated features
- elimination of insignificant features with one-way ANOVA F-test for classification and correlation coefficient for regression

1) Redundant features: For each categorical input feature, the number of categories is measured. Two extreme cases are eliminated based on the number of unique values:

- Constant feature: single category features, eg., Machine No., Nationality, etc.,
- Feature with the number of categories equal to the number of samples, eg., Name, Email Id, etc.,

In both cases, the information gain is zero and hence the features are removed.

2) Correlation Threshold: In this step, a one-dimensional correlation among the input features is calculated. Pearson's correlation [19] coefficient as shown in (1) is calculated for all input features. Features are merged based on their correlation coefficient. Features with a correlation coefficient close to 1 would provide similar information for modeling and can thus be removed to save the computational power. Equation 1 shows the correlation between two features denoted as x and y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where, x_i, y_i are the i^{th} elements for features x and y.

3) Analysis of Variance: After the elimination of insignificant features from steps 1 and 2, the rest of the features are analyzed based on a one-way ANOVA test for a classification task. Each feature is divided into subgroups corresponding to its target value. The one-way analysis of variance is carried out with these subgroups to find the F-value, as described below.

$$\begin{aligned} SSR &= \sum_{i=i}^{k} n_i(\bar{X}_o - \bar{X}_i)^2 \\ SSE &= \sum_{i=i}^{k}\sum_{j}^{n_i}(\bar{X}_i - X_{ij})^2 \\ SST &= SSR + SSE \\ df_r &= k - 1 \\ df_e &= \sum_{i=1}^{k} n_i - k \\ MSR &= \frac{SSR}{df_r} \\ MSE &= \frac{SSE}{df_e} \\ F\_value &= \frac{MSR}{MSE} \end{aligned} \tag{2}$$

where k is the number of groups, n_i is the number of samples in group i, $\bar{X}_o$ is the global mean of all samples of the features and $\bar{X}_i$ is the mean of samples in subgroup i, df_r is regression degree of freedom and df_e is error degree of freedom. MSR is regression mean square, MSE is error mean square, SSR represents the regression sum of squares, SSE is the error sum of squares and SST is the total sum of squares.

For the regression datasets, F-value is calculated via a univariate linear regression test. Once the F-values are available for all features, they are ranked based on $p\text{-}values$ as suggested by Charles Poole [9]. Experimentally, it is found that a relative comparison of $p\text{-}values$ is a better evaluation of the significance of the feature. The features are arranged in decreasing order of $\text{-}log(p\text{-}values)$ and the maximum drop of $\text{-}log(p\text{-}value)$ over the features is computed by calculating their second-order gradients. The maximum drop in the $\text{-}log(p\text{-}value)$ is considered a threshold for that particular dataset. Features with $\text{-}log(p\text{-}value)$ less than this threshold are considered insignificant. In [9], a $p\text{-}value$ of 0.05 is considered an appropriate threshold value for selecting the insignificant features. Experimentally, it is found that a few datasets have many features with $p\text{-}values > 0.05$. To overcome this situation where many features could be removed because of a static threshold, the second-order gradient method

 ISBN: 978-1-61208-976-8

is used. By using a second-order gradient, the maximum drop of the $-log(p\text{-}value)$ can be detected. The advantage of this method is the threshold for the $p\text{-}value$ and it is set dynamically. The least important features are identified with the threshold as mentioned below:

$$threshold = \arg\max \left\{ \nabla \left(- \log \left(p\_values_{sorted} \right) \right) \right\} \quad (3)$$
$$features = features_{(p\_value < threshold)}$$

Based on the above three methods, the most important features are selected for later processing.

B. Sampling

The quality of a sampling method can be accessed based on the divergence of a sampled dataset with the original distribution. As mentioned in Section II, the most widely used sampling techniques are random sampling, stratified sampling, and hand-pick sampling [15]. Stratified sampling is the basis for our proposed approach, Bin-Based sampling.

1) Stratified Sampling: Stratified sampling [14] by Cochran is a well-known sampling technique that closely resembles the original distribution statistics. In stratified sampling, a population is divided into sub-populations *strata* followed by random sampling from these *strata*. The proportionate allocation of sampling steps for the creation of strata is summarized in Algorithm 1. These sub-populations are formed based on the nested grouping of columns. One of the disadvantages of stratified sampling is its non-realistic runtime which makes it difficult for real-time applications. Therefore, we propose a new sampling technique, Bin-Based sampling. Two major advantages of bin-based sampling are:

- faster than stratified sampling
- preserves original distribution better than random sampling

2) Bin-based sampling: The motivation behind bin-based sampling is to reduce the time complexity while maintaining the original population statistics. To achieve this, input features are divided into different bins based on their distribution. After the binning process, random samples are drawn from each bin for every feature. A union set of sample collection from every feature is now the bin-based-sampled population. Figure 2 shows the histogram of a feature before and after sampling. We can see that the distribution is preserved with Bin-Based sampling and it simplifies the probability of choosing a sample by lowering the number of possibilities. The joint conditional probabilities in stratified sampling are simplified into the conditional probabilities of each feature, as mentioned below.

$$P_{random}(s) = \frac{1}{N}$$
$$P_{bin-based}(s) = \frac{P(s|b_i)P(b_i)}{P(b_i|s)} \quad (4)$$
$$P_{bin-based}(s) = P(s|b_i) = \frac{1}{size(b_i)}$$

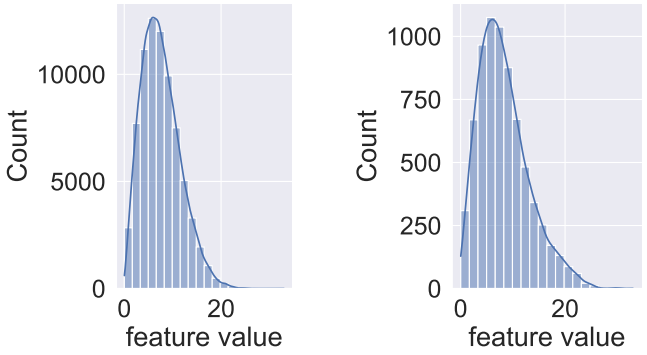

Fig. 2. Bin-based Sampling: Comparison of distribution for a feature before and after Bin-based sampling. (left: original distribution; right: sampled distribution)

where N denotes the number of samples, s is the sample point and b_i denotes the i^{th} bin. The pseudo-algorithm for Bin-Based sampling is provided in Algorithm 2.

Algorithm 1: Stratified sampling

```
Function Subsets(Data):
    for category/bin in feature_i from Data do
        if num_features in Data > 1 then
            Strata ← Subsets(Data.drop(feature_i))
        else
            Strata ← category/bin
    return Strata
Strata ← Subsets(Data)
for Stratum in Strata do
    stratumSamples ← RandomSample(stratum)
StratifiedSample ← Concat(stratumSamples)
Result: StratifiedSample
```

Algorithm 2: Binbased Sampling

```
for Feature_i in Features do
    if Feature_i is Categorical then
        for Category in Categories_Feature_i do
            Sample ← RandomSample(Category)
        end
        featureSample ← Concat(Sample)
    else
        Bins ← Discretize(Feature_i)
        for Bin in Bins do
            Sample ← RandomSample(Bin)
        end
        featureSample ← Concat(Sample)
    end
end
BinbasedSample ← Concat(featureSample)
Result: Binbased Sample
```

3) Sampling size: An optimal sampling size should ensure that the information loss is minimum. Cochran [14] has stated

an optimal size for sampled population based on the size of the population.

$$n_o = \frac{Z^2 p(1-p)}{e^2} \tag{5}$$

where e is desired level of precision (i.e., margin or error), p is the estimated proportion of the population, and Z is the z-score distribution value, defaulting to 0.475 for 95% distribution. For Bin-Based sampling, a sample size with a z-score distribution value of 0.475 is chosen, as suggested by Cochran [14].

C. Target Discretization

With the help of target discretization, a numerical output feature can be converted into categorical values, thereby transforming a regression task into a classification task. Based on the baseline regression model, a regression task will be converted to a classification task if the regression r2-score metric value is significantly low or unacceptable. The prediction of categorical values has less degree of freedom than the prediction of numerical values. Taking advantage of this fact, a classification analysis with AutoML might give a reasonable classification accuracy rather than concluding the datasets as not suitable for analysis. Here, each data point in the continuous domain is converted into a discrete class domain. Different types of target discretization methods can be considered based on domain expertise. As an automated solution, we have considered the discretization of the target variable based on its z-score values.

D. Feature Engineering

Feature engineering is the process of generating new features or transforming features from the existing set of features using domain knowledge. Feature engineering is performed to leverage the performance of the ML model. Since domain knowledge is not always available, we suggest an automated way to feature engineering - Hybrid Feature engineering, inspired by Cognito [1] and ExploreKit [2]. They are briefly described below.

1) Cognito: As mentioned in Section II, Cognito generates new features by performing unary and binary operations on the input features. As the number of input features and transform operators increases, the number of newly generated features increases exponentially in the order of $O(f.d^{k+1})$ where f is the number of features, d is the number of combinations and k is the number of transforms. For a feature D and transforms τ_1 and τ_2, $\tau_1(D)$, $\tau_2(D)$, $\tau_1\tau_2(D)$ are generated. These generated features have to be pruned to reduce the complexity of the problem. Feature selection is done on the generated features using information gain as a proxy measure of accuracy.

2) ExploreKit: ExploreKit [2] takes a similar approach towards feature engineering. Together with the unary and binary operators, ExploreKit considers higher-order operators. Among all the generated features, each feature is added to the dataset, and the rank of the feature is determined with the Ranking Model. The Ranking Model ranks the newly generated features based on either accuracy score (classification) or r2-score (regression). Low-rank features are removed and higher rank features are added to the original dataset so that more information can be extracted.

3) Hybrid Feature Engineering: Inspired by Cognito and ExploreKit, we propose a new Hybrid Feature engineering approach for feature engineering, which can be implemented in real-time with a modified ranking algorithm and additional usage of the Bin-Based sampling method. In Hybrid feature engineering, new features are generated with a single transformation on a feature or features at a time. This transformation are either a unary or a binary operator. For a feature D and unary transforms τ, $\tau(D)$ feature is generated. For features D_1, D_2 and binary transform τ, $\tau(D_1, D_2)$ feature is generated. These features are subjected to feature selection as described in Section III-A instead of using the Ranking Model directly. After the elimination of features with feature selection, most of the insignificant features are eliminated and we are left with a relatively less amount of features. These features are then ranked with the Ranking Model. High-ranked features are selected and added to the original dataset.

Consider $F = f_1, f_2, ..., f_n$ is a set of features, $T = \tau_1, \tau_2, ..., \tau_n$ is a set of transform functions and $F' = FXT$ denotes the set of new generated features. With the help of the Feature Selection technique, the most significant features, $I \subset FXT$ can be selected from the generated features. After Feature Selection, the set I is fine-tuned with Ranking Model R. (c.f. Algorithm 3, Algorithm 4)

$$I \leftarrow R(FXT) \tag{6}$$

Algorithm 3: Hybrid Feature Engineering

```
for Operator in Unary/Binary Operators do
    for Feature in Numerical-Features do
        NewFeatures ← Operator(Feature)
    end
    allNewFeatures_operator ←
     FeatureSelection(NewFeatures)
end
allNewFeatures = RankingModel(allNewFeatures)
Result: Generated Features
```

Algorithm 4: Ranking Model

```
thresholdf = baseModel(Dataset)
for i = 0 to allNewFeatures do
    Dataset.append(allNewFeatures[i])
    RankedFeatures = []
    featureScore = baseModel(Dataset)
    if featureScore ≤ thresholdf then
        continue
    else
        RankedFeatures.append(allNewFeatures[i])
    end
    return RankedFeatures
end
Result: Ranked features
```

 ISBN: 978-1-61208-976-8

IV. Experiments and Results

This section describes various experiments that are conducted together and their validation results. The goal of this validation study is to ensure that the proposed methods have a positive influence on the datasets. Validation is done on OpenML datasets [18]. RandomForest model is used as a baseline model. The proposed auto-preprocessing pipeline is also benchmarked against top-performing AutoML libraries [17], namely AutoSklearn, Autogluon, and H2O. The results are consistently compared with (w) and without (w/o) the auto-preprocessing libraries. Figure 1 illustrates the flow of the preprocessing pipeline.

A. Experimental Setup

For an individual experiment, a RandomForest model with 30 estimators was chosen for both classification and regression tasks. K-fold cross-validation was used to benchmark the results. All experiments are conducted on a Linux Virtual Machine with 16 GB of RAM and 4 cores. Experiments are conducted without using `dask` multiprocessing for AutoML libraries.

B. Datasets

OpenML datasets were used for benchmarking the results [6]. A comparative study shows that Hybrid Feature Engineering (HFE) performs better than modeling without HFE for around 35% of cases, no change in performance was observed for the rest of the datasets. This trend can be observed for both classification and regression tasks. It has to be noted that an increase in performance can be expected only if we have new features after the pruning method as explained in Section III. Only 35% of the OpenML datasets reported new features after the pruning step. Results of these datasets are provided in Tables II - VIII.

C. Feature selection

A comparative study was conducted to compare the performance with and without feature selection. It can be seen that for all datasets performance remains the same after the elimination of less significant features, as described in Section III-A. Reduction in the training time is not very significant for the baseline model, as the model has only 30 estimators and the size of the datasets is comparatively less. For MNIST, a total of 64 features out of 784 features were eliminated maintaining the same accuracy. Feature selection is the first step in the preprocessing pipeline. The results of the analysis are summarized in Tables II and III.

D. Bin-Based sampling

Bin-based sampling is used to reduce computation time for the baseline model used for feature engineering. Stratified sampling is known to sample a good representation from a population but at the cost of computation time with a time complexity O(n^2) where *n* is the number of input features. Random Sampling, on the other hand, has the complexity of O(1) and Bin-Based sampling, as explained in Section III-B O($\frac{n}{3}$). The sampling technique is validated using Kullback-Leibler (KL) divergence (7) considering the original distribution as the reference distribution [16].

$$D_{KL}(P\|Q) = \int_{-\inf}^{\inf} P(x) log\left(\frac{P(x)}{Q(x)}\right) dx \qquad (7)$$

where $P(x)$ is the original distribution and $Q(x)$ is the sampled distribution. We can observe that KL-divergence for Bin-Based sampling is comparatively much lower than random sampling. Stratified sampling has the least KL-divergence, but 10 times the computation time as can be inferred from Table IV. These experiments are conducted on a Linux VM with 256 GB of RAM. Experiments also revealed that Bin-Based sampling failed to perform well for the datasets of smaller sizes. The reason for this is that performing a binning operation and extracting random samples from each bin might cause a loss of information. Therefore, for such datasets of small size, sampling is not useful.

E. Hybrid Feature Engineering

Significant improvement in performance is achieved for a few OpenML datasets. These are summarized in Tables V and VI. A performance improvement for 35% of datasets is observed when testing over 50+ OpenML datasets. All tests are performed on a cross-validation split. It should be noted that performance improvement is achieved only if new significant features are generated after the pruning step. The performance with HFE is higher or similar, in no case did we encounter a decrease in performance.

F. Overall Pipeline

An auto preprocessing pipeline, as shown in Figure 1, is used together with AutoML libraries mainly AutoGluon, AutoSklearn, and H2O for benchmarking. Significant improvements compared to the baseline model are not achieved with AutoML libraries for all datasets shown in Tables V and VI. The results with AutoML libraries are shown in Tables VII and VIII. The stopping criteria for training of AutoML-libraries are set with the runtime limit. The benchmarking for AutoML libraries is done on a single core. The runtime across all AutoML libraries is set to 10 minutes and the results are 4-fold cross-validated. Overall, the combination of auto-preprocessing and AutoML libraries performed better or was similar to "only AutoML libraries". As mentioned in Table I, AutoML libraries do not consider feature engineering in their preprocessing step.

V. Conclusion

A significant amount of recent work in the field of automated-Machine Learning is being done, but the same has not been the case for data preprocessing. This paper reviews and suggests some advanced preprocessing steps that can either be used individually or combined as a pipeline. Although performance improvements cannot be ensured for all datasets, datasets that have inter-feature dependency can be observed to perform better. For example, length and width

 ISBN: 978-1-61208-976-8

TABLE II
VALIDATION OF FEATURE SELECTION TECHNIQUE FOR CLASSIFICATION TASK

OpenML dataset	Accuracy w/o feature selection	Accuracy with feature selection	Number of features removed	Difference in accuracy
11	0.607	0.607	3	0
54	0.753	0.753	0	0
188	0.579	0.579	0	0
333	0.908	0.908	3	0
335	0.977	0.977	2	0
470	0.661	0.661	4	0
1459	0.588	0.588	0	0
1461	0.692	0.692	2	0
23381	0.560	0.560	5	0
amazon-employee-access	0.943	0.943	3	0
australian	0.857	0.857	2	0
bank-marketing	0.692	0.692	2	0
credit-g	0.761	0.761	2	0
sylvine	0.941	0.941	7	0

TABLE III
VALIDATION OF FEATURE SELECTION TECHNIQUE FOR REGRESSION TASK

OpenML dataset	r2-score w/o feature selection	r2-score with feature selection	Number of features removed	Difference in r2-score
537	0.484	0.484	0	0
495	0.616	0.616	5	0
344	0.999	0.999	2	0
215	0.948	0.948	1	0
189	0.579	0.579	0	0
507	0.391	0.390	0	0

TABLE IV
SAMPLING COMPARISON ON OPENML DATASETS CALCULATED OVER 100 TRIALS

OpenML dataset	Mean of KL-divergence Bin-Based sampling	Mean of KL-divergence Stratified sampling	Mean of KL-divergence Random sampling	Time (in sec) Bin-Based sampling	Time (in sec) Stratified sampling
183	0.017	0.173	0.057	0.359	5.230
223	0.067	0.079	0	0.273	7.600
287	0.076	0.356	0.027	0.399	4.807
307	0.0	0.097	0.006	0.214	7.572
528	0.0	0.0215	0.0	0.054	0.489
537	0.190	0.886	1.160	2.052	133.939
550	0.0	0.011	0.004	0.302	0.738
Amazon-employee-access	0.019	0.460	0.753	0.466	2.112
Blood-transfusion	0.065	0.002	0.001	0.062	0.069
Phoneme	0.0	0.168	0.143	0.580	1.383

of a workpiece can be combined to form a new feature "area of the workpiece", which can have a significant impact on the ML-based model. The proposed method does it without domain knowledge in an automated manner. This paper also introduces a new sampling method that can be used for general application as well as for ML-based modeling. We used the Bin-Based sampling method during the Feature Engineering step to generate new features and select them using a Ranking Model. Usage of sampled data for Feature Engineering has significantly reduced the preprocessing time. It can be concluded that a significant performance improvement of around 4-7% is observed for the analysis conducted with the baseline model on OpenML datasets. For the same set of datasets, a marginal improvement was observed for analysis with the AutoML libraries. The proposed pipeline is currently not parallelized. Parallelization can significantly reduce the time for feature engineering and this we would like to focus on in our future work.

REFERENCES

[1] U. Khurana, D. Turaga, H. Samulowitz and S. Parthasrathy, "Cognito: Automated Feature Engineering for Supervised Learning." *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016: pp. 1304-1307.

[2] G. Katz, E. C. R. Shin and D. Song, "ExploreKit: Automatic Feature Generation and Selection," *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 979-984, doi: 10.1109/ICDM.2016.0123.

[3] S. Galhotra, U. Khurana, O. Hassanzadeh, K. Srinivas, H. Samulowitz and M. Qi, "Automated Feature Enhancement for Predictive Modeling using External Knowledge," *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 1094-1097, doi: 10.1109/ICDMW.2019.00161.

[4] H. T. Lam, T. N. Minh, M. Sinn, B. Buesser, and M. Wistuba, "Neural Feature Learning From Relational Database." *arXiv: Artificial Intelligence*, 2018.

 ISBN: 978-1-61208-976-8

TABLE V
HYBRID FEATURE ENGINEERING FOR CLASSIFICATION DATASETS WITH BASELINE MODEL

OpenML datasets	Number of features	Number of classes	Accuracy before	Accuracy after	Percentage Gain	New features
188	14	5	0.466	**0.506**	8.386	2
1461	7	2	0.692	**0.718**	4.748	2
1459	7	10	0.588	**0.635**	7.952	1
54	18	4	0.753	**0.759**	0.786	2

TABLE VI
HYBRID FEATURE ENGINEERING FOR REGRESSION DATASETS WITH BASELINE MODEL

OpenML datasets	Number of features	r2-score before	r2-score after	Percentage Gain	New features
189	8	0.579	**0.615**	6.227	1
507	6	0.390	**0.411**	5.361	1
537	8	0.484	**0.494**	2.000	1
495	13	0.616	**0.632**	2.700	2

TABLE VII
OVERALL PREPROCESSING PIPELINE PERFORMANCE COMPARISON WITH AUTOML LIBRARIES (CLASSIFICATION - ACCURACY)

OpenML datasets	AutoGluon		AutoSklearn		H2O		RandomForest	
	w/o	w	w/o	w	w/o	w	w/o	w
188	0.728	0.726	0.674	0.696	0.717	0.739	0.466	0.506
1461	0.914	0.914	0.906	0.906	0.907	0.907	0.692	0.718
1459	0.815	0.82	0.919	0.919	0.922	0.927	0.588	0.635
54	0.858	0.857	0.839	0.839	0.707	0.708	0.753	0.759

TABLE VIII
OVERALL PREPROCESSING PIPELINE PERFORMANCE COMPARISON WITH AUTOML LIBRARIES (REGRESSION - R2-SCORE)

OpenML datasets	AutoGluon		AutoSklearn		H2O		RandomForest	
	w/o	w	w/o	w	w/o	w	w/o	w
189	0.913	0.913	0.902	0.903	0.913	0.918	0.579	0.615
507	0.731	0.741	0.753	0.753	0.762	0.761	0.390	0.411
537	0.815	0.821	0.862	0.865	0.861	0.869	0.484	0.494
495	0.496	0.495	0.494	0.494	0.441	0.442	0.616	0.632

[5] J. M. Kanter and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors." *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015: pp. 1-10.

[6] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum and F. Hutter, "Efficient and Robust Automated Machine Learning." *NIPS 2015*.

[7] E. LeDell and S. Poirier, "H2O AutoML: Scalable Automatic Machine Learning". *7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020.

[8] N. Erickson et al., "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data." *ArXiv abs/2003.06505*, 2020.

[9] C. Poole, "Low P-values or narrow confidence intervals: which are more durable?" *Epidemiology* 12, 2001: pp. 291-294.

[10] T. Terano, H Liu and L. P. Arbee, "Chen Knowledge Discovery and Data Mining." *4th Pacific-Asia Conference, PAKDD 2000*, Kyoto Japan, 2000.

[11] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation." *J. Mach. Learn. Res. 11*, 2010, pp. 171–234.

[12] C. Jie, L. Jiawei, W. Shulin and Y. Sheng, "Feature selection in machine learning: A new perspective." *Neurocomputing* 300, 2018: pp. 70-79.

[13] N. O. F. Elssied, O. Ibrahim and A. H. Osman, "A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification." *Research Journal of Applied Sciences, Engineering and Technology 7*, 2014: pp. 625-638.

[14] Cochran William, Sampling Techniques, 3rd edition, *John Wiley and Sons*, 1978.

[15] J. A. R. Rojas, M. B. Kery, S Rosenthal, and A. Dey, "Sampling techniques to improve big data exploration." 2017 *IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)*, 2017: pp. 26-35.

[16] J. M. James, "Kullback-Leibler Divergence". *International Encyclopedia of Statistical Science*, 2011.

[17] P. Gijsbers, E. LeDell, J. K. Thomas, S. Poirier, B. Bischl, and J. Vanschoren, "An Open Source AutoML Benchmark." *ArXiv abs/1907.00909*, 2019.

[18] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: networked science in machine learning." *SIGKDD Explorations 15(2)*, 2013, pp. 49-60.

[19] Kirch Wilhelm, "Pearson's Correlation Coefficient." *Encyclopedia of Public Health*, 2008. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-5614-7_2569 .

 ISBN: 978-1-61208-976-8

Valuing Built Heritage through the Promotion of Oral Heritage via Participation in the Digital Age

Feedback from the Stories of the Old City of Kairouan

Khaoula Stiti
BATir dept. - Université Libre de Bruxelles
Brussels, Belgium
National School of Architecture and Urban Planning – University of Carthage
Tunis, Tunisia
e-mail: Khaoula.stiti@ulb.be

Safa Achour
National School of Architecture and Urban Planning – University of Carthage
Tunis, Tunisia
e-mail: safa.achouryounsi@enau.ucar.tn

Samia Ben Rajeb
BATir dept. - Université Libre de Bruxelles
Brussels, Belgium
e-mail: samia.ben.rajeb@ulb.be

***Abstract*— As part of the tools of the recent global boom in the democratization of knowledge, Information and Communication Technologies (ICT) have been giving growing support to participation in cultural heritage. This paper presents "*Ahkili Aliha*", an event during which tribunes of speech were open to the inhabitants of Kairouan, Tunisia, serving the (re)promotion of the patrimonial site of Kairouan through the promotion of oral heritage and collective memory. Based on the *Ahkili Aliha* event case study, the main objective of this research is to examine the relevance and contribution of participation and ICT for the promotion of oral heritage and consequently the promotion of built heritage of patrimonial sites. The study presents critical analyses of the linearity of the patrimonialization process in old Tunisian cities and considers the role of participation and ICT in resuming paused or incomplete patrimonialization processes.**

Keywords- participation; ICT; oral heritage; built heritage; patrimonialization; collaborative action research.

I. INTRODUCTION

Two perceptions of the heritage seem to be often in opposition: while researchers perceive the heritage as a testimony of the history and a necessary element to preserve the memory of the past, the people perceive the heritage as buildings and territories to be lived in according to living arrangements today [1]. At the same time, another opposition is existing between tangible and intangible heritage. Eichler [2] emphasizes the existence of patterns of neo-colonialism that may subject communities to complementary power asymmetries that jeopardize their authoritative voices throughout intangible cultural heritage practices: neo-colonial relations become apparent in a variety of ways, encompassing written versus oral heritage, material versus intangible heritage, disadvantaged regions under the UNESCO umbrella and ultimately competing State alliances played out to the detriment of the Global South. The main objective of this article is to examine (1) the capacity of participation and Information and Communication Technologies (ICT) to promote oral heritage, and (2) the capacity of oral heritage promotion on built heritage. We focus on how intervening in collaborative action research makes possible and activates the feeling of belonging of participants, allowing them to retrace the history of the city and better promote the oral and built heritage. For this purpose, the article is based on feedback from the collaborative action-research project *Ahkili Aliha*. This project aimed to involve young people in the promotion of the historical center of their city by creating an immersive journey through stories told by the inhabitants retracing the memory of places and buildings. The specificity of this project was to promote the patrimonial site of the old city of Kairouan through activities carried out by youth and non-experts in cultural heritage and history while involving experts and researchers in these fields. The aim was to maintain links of dialogues without hierarchal distinction between institutional knowledge and non-institutional knowledge and between tangible heritage and intangible heritage, in this case, oral heritage. The article first describes the frameworks we used to carry out our study: the context of the feedback project *Ahkili Aliha*, its approach: the collaborative action research, and the theoretical framework: the process of patrimonialization of Tunisian old cities. Then, it exposes the research design through a presentation of the research problem, the research methods, and the case study of *Ahkili Aliha*. The overview of the results allowed us to validate the role of participation and ICT in heritage promotion through the examination of the different components of the project *Ahkili Aliha*. The study presents critical analyses of the linearity of the patrimonialization process in old Tunisian cities and considers the role of participation and ICT in resuming paused or incomplete patrimonialization processes.

 ISBN: 978-1-61208-976-8

II. RESEARCH FRAMEWORKS

A. *Research context:* Ahkili Aliha *project*

The researchers of Université Libre de Bruxelles, Belgium (ULB) and the members of Edifices & Mémoires, Tunisia (E&M) defined the *Ahkili Aliha* project in line with mobilizing young people to fight against the destruction of cultural heritage. *Ahkili Aliha* means in Arabic "tell me about it/her". In this context, *Ahki* can also be translated as narrate, tale and yarn. The targeted objectives were to enrich a collective memory around multiple and shared values and to contribute to the feeling of common belonging: living together. Kairouan, known as one of the oldest UNESCO sites in Tunisia and referred to as the fourth holy (or sacred) city of Islam and the first holy city in North Africa, has witnessed several activities related to violent extremism in recent years. The Medina of Kairouan has problems with security and delinquency and several young people have been radicalized. To tackle this issue, *Ahkili Aliha* project aimed to promote a sense of identity and belonging in youth in their city. The funding of *Ahkili Aliha* was obtained from the "Prevent Violent Extremism" fund from the United Nations, co-funded by the Government of Canada.

The vision of the project was to help ward off violence and extremism is to help youth speak up and/or act. To do this, the research over nine months (April to November 2019) had brought together different partners:

- Key civil society organization: Edifices & Mémoires (E&M), a Tunisian nongovernmental organization working for the preservation and promotion of the Tunisian architectural and urban heritage through communication between scientific research and citizen action. It aims to bring together several disciplines related to the appropriation of heritage to go beyond simplistic museification in favor of a global, contextualized, and multidisciplinary reflection on heritage and the awareness of collective memory. E&M considers that the challenge is to move from a "frozen" heritage to a heritage designed and lived by all.
- Scientific partner: research unit AIA Architectural engineering, BATir department, Ecole Polytechnique; Université libre de Bruxelles. The unit is specialized in participatory approaches and the study of citizen actions in favor of collective intelligence.
- Institutional partner: National Heritage Institute, Tunisia. The immediate partner was the regional office of the National Heritage Institute in Kairouan.
- Local civil society organizations: the main local civil society organizations are: We Love Kairouan, Junior Chamber International of Kairouan, UNESCO, ISESCO, ALECSO Club of Kairouan. These organizations helped reach a big number of the population, especially young people, for the implementation of the project.

The complete organigramme of *Ahkili Aliha* is presented in Figure 1 below.

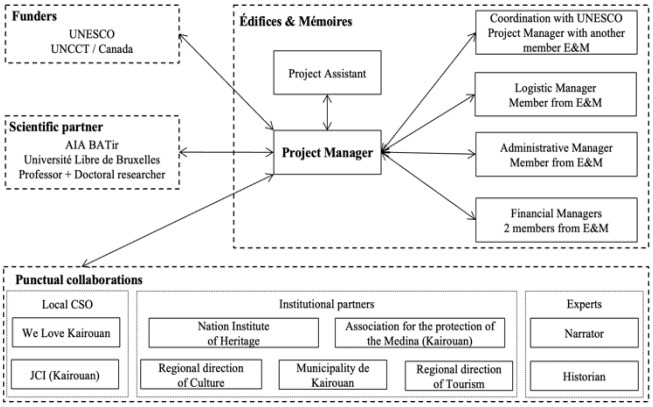

Figure 1. Ahkili Aliha organizational chart

In previous work "in press" [3], we classified participation actors into three groups: actors by knowing, actors by knowledge, and actors by action. In *Ahkili Aliha* project, actors by knowing are represented in the local civil society organizations as well as inhabitants who participated in the project. Actors by knowledge are represented in the scientific partner, the experts, as well as the managers from E&M since they are all architects and experts in built heritage. Actors by action are represented in the funders and the institutional partners in Kairouan. The project manager played an important and crucial role in the project implementation. The financial, logistic, and administrative managing was guaranteed in collaboration with the different officers from E&M. Born and raised in the city of Kairouan, the project manager had ease to connect with the local civil society organizations for the punctual collaboration. The scientific partners cooperated with the project manager to guarantee the scientific application of the participatory approach followed in *Ahkili Aliha*. A regular contact with funders was established between the project manager and the administrative representative of UNESCO and the "Prevent Violent Extremism" fund. Since the project manager was also an architect and a doctoral researcher in ICT use in heritage promotion, she was an actor by knowing and an actor by knowledge at the same time. Collaborative action-research was the approach followed in *Ahkili Aliha* and the framework for the interactions between actors.

B. Ahkili Aliha *as action-research project*

Training in the safeguarding and in the promotion of cultural heritage and encouraging and supporting creativity among youth was among the aims of this action-research project. Action research is a proactive strategy in which research has political and social relevance. Action research in

 ISBN: 978-1-61208-976-8

architecture and design is a form of learning through doing/making, which chimes both with Paulo Friere's conceptions of learning during action [4], and importantly with a core tenet of participatory design, mutual learning [5]. In action research, it is important from an epistemological point of view to clearly identify the difference between the research question targeted by researchers and the action expected by the people. In *Ahkili Aliha*, the researchers aimed to validate the role of participation and ICT in valuing oral heritage to promote built heritage, while the action fulfilled by the inhabitants, elder people and youth especially, was to organize the event of *Ahkili Aliha*, as a day where both oral heritage and built heritage from aside, and both institutional knowledge and non-institutional knowledge are celebrated.

C. Theoretical framework: Patrimonialization of Tunisian old cities

In this section, we build a theoretical framework for patrimonialization based on the works of Emmanuel Amougou [6] and Zeineb Youssef [7] to discuss later in this paper the roles that participation and ICT can play in the patrimonialization process. In the book "The patrimonial question", the process of patrimonialization creates links, requirements, and implications according to Amougou. Among these elements, we cite (1) the social global links (the emergence of the question of heritage and its (re)definition, (2) the institutionalization (the mechanisms of diffusion of the patrimonial legitime values), (3) the professionalization (confirmation invention of agents on object, knowledge speech, practices and techniques of legitimation), (4) new social links appearing to redefine objects from new social challenges) and (5) the practical traductions (applications of different dispositions such us restoring, renovation, etc.).

Youssef [7] emphasizes that the process of patrimonialization is linear and composed of phases sub-composed of steps. Thus, as said by Youssef, the construction of the patrimonial profile for Tunisian medinas is possible within 6 different scenarios according to the steps they are composed of. The process of patrimonialization is composed of 3 phases: identification, conservation, and exploitation. The first phase of identification is composed of awareness and selection. The second phase of conservation is composed of protection and conservation. The final phase of exploitation is composed of exhibition and valorization. Different scenarios can result from the presence and/or the absence of steps in each context. The following table explains each scenario according to the patrimonialization steps achieved.

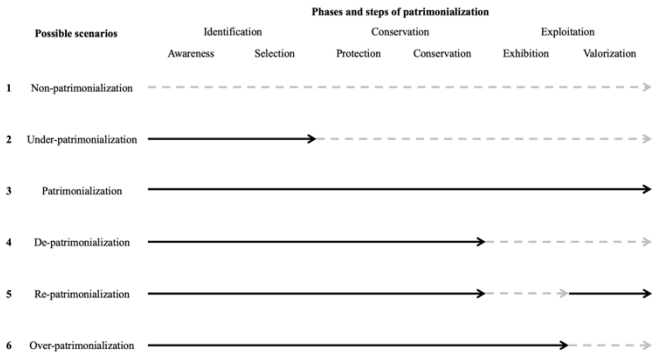

Figure 2. Phases and steps of patrimonialization adapted from Youssef [7].

III. RESEARCH DESIGN

A. Research problem

In the *Ahkili Aliha* case study that we analyze in the following pages, we shall examine the capacity of participation and ICT to promote oral heritage, and therefore, the promotion of built heritage. We focus on how intervening in collaborative action research makes possible and activates the feeling of belonging of participants, allowing them to retrace the history of the city and better promote the oral and built heritage.

B. Research methods

The research design is based on two techniques involving collecting qualitative data and adapted to action-research projects. While the first part consists of the research design aims to reconstruct the *Ahkili Aliha* project process through documents analysis, the second part is based on participant observation carried out during the first participatory workshop and meetings with the E&M team. Since participant observation didn't take place during the entire project, it was completed by the documents analysis technique. Thus, it can be said that this study uses grounded theory methods composed of two different techniques to achieve comprehensive and authentic data collection.

1) In-Depth documents analysis

Document analysis is often used in combination with other qualitative research methods as a means of triangulation 'the combination of methodologies in the study of the same phenomenon [8]. Drawing upon multiple (at least two) sources of evidence; that is, to seek convergence and corroboration using different data sources and methods. Apart from documents, such sources include interviews, participant or non-participant observation, and physical artifacts [9]. The documents that we analyze in our study are produced before, during, and after the project, mainly project submission documents, periodic reports to funders, and emails between the people involved in the project.

 ISBN: 978-1-61208-976-8

2) Participant-observation

Participant observation can be used to explore the entire range of themes and topics that qualitative research methods are generally used for and may be particularly useful whenever the aim is to understand a phenomenon or a setting from the perspective of those who live, experience, and/ or are affected by it. It is also a research approach that may be exploited in participatory action research [10]. We were able to observe the interactions between the people involved in the project during the project and after that. In the next section, we describe the two participatory workshops which took place in the beginning and in the middle of the *Ahkili Aliha* project, since they were the major occasion where different actors had the opportunity to meet and design together.

C. Choice of components of Ahkili Aliha *Project*

The phases of the collaborative methodology of the *Ahkili Aliha* project had been specified to ensure the participation of the partners in all the phases of the project. Table I details each component (first column) by specifying the different steps that have been given and prepared (second column) and the activities realized (third column) with the partners involved (fourth column).

TABLE I. COMPONENTS OF THE *AHKILI ALIHA* PROJECT

Components	Steps	Activites	Partners included
I. Promotion of the collective memory of a place, *the route*	**Taking contact**	- Town and neighbourhood visit - Identification of historical buildings - Selection of the lead narrator	- Édifices & Mémoires - Local CSO
	Participatory workshop	- Study of selected buildings to identify the main elements of the route: history, architectural value, personal value, choice of place - Identification of locations (including the location of the central stage) according to the potential of the place to convert it into a stage, the proximity of the different places with the length of the route, the accessibility - Checking the coherence between the selected buildings and the main narrative told on the urban scene - Identification of related activities - Route design	- Édifices & Mémoires - Lead narrator - Local CSO - Scientific partner
	Realization	- Dissemination of the event and its publication - Preparation of related activities (local products booths) - Preparation of guided tours - Preparation of music concert by local musicians at the end of the route	- Cultural Professionals (artists, experts in cultural management and in mediation) - Édifices & Mémoires - Local CSO - Scientific partner
	Simulation	- Design an urban scene in the city (also starting point for the guided tours)	All partners
II. Promote the individual memory of a place, *the box*	**Taking contact & workshop**	- Meet the young people interested in the project through local CSO - Select Profiles - Identify the stories to capture in the box - Define the place where the box is set up	- Professionals in video capturing and processing - Édifices & Mémoires - Lead narrator - Local CSO - Scientific partner
	Training	- Prepare young people to conduct an interview with the people to be interviewed in the box	
	Box design & installation	- Choosing the van where the filming will take place - Customizing the van with the parameters of visibility, its identity image, its intimacy as the filming place and the security of the equipment contained therein	
	Concretisation	- Filming and capturing - Processing the captured data	- Professionals in video capturing and processing - Édifices & Mémoires - Narrators - Local CSO - Scientific partner
	Simulation	- Test the feasibility of the projection and the equipment at the selected locations along the route	
III. *Ahkili Aliha* narration & route event		- Urban storytelling event: guided tours provided by the young people who participated in the workshops: screening of the catches, exhibition of local products, discovery of the built heritage of the city - Musical concert on stage	All partners

IV. RESEARCH RESULTS

A. Participation to promote the oral heritage

1) Participation to promote the collective memory of a place, the Route

The participatory workshop that took place at the beginning of the project allowed the participants to express their relationships with the heritage of their city as well as engage in participatory reflections to choose together the route and the delimitation of the area of activities. This co-design of the event encourages formal and informal moments of exchange, by mixing all the partners, which allowed a better appropriation of the project by the participants. The co-creation of the route allowed the reshaping of the collective memory of the city that the participants have in common.

Figure 3. The partners during the participatory workshop.

2) Participation to promote the individual memories of a place, Stories from the Box

The box installed open to the public served to capture the stories of the city through the testimonies of its inhabitants. This box was a van equipped with cameras, microphones, and light. The van was rented and decorated on both sides and from behind with a catchphrase written in Tunisian dialect to attract the attention of passers-by. The project took the same design of the van's skin for the banner to keep the same visual identity. The banner was fixed on the rampart of the old city in the Martyrs Square where the billboards of the major cultural events and citizen actions of the city are fixed. It is also used to warn passers-by, when the van is not parked, of the date and time of the opening of the studio. The van was parked all week of registration in the same square, which gives access to the Medina. It was the liveliest square with the most frequented cafés by most of the people who live or have lived in Kairouan and who present the project's target for the interviews. In front of this square, there is a commercial district and the cultural complex of the city which are frequented rather by young people. The studio is open every day at the Martyrs Square from 6pm to 10pm to avoid the daytime heat and to have the presence of the inhabitants who frequent the square to drink a coffee, meet their friends or take a walk in the old city. In the mornings, the van recording team went to people who could not move to the Martyrs Square and with whom they have made appointments in advance to interview them. Therefore, this was an opportunity to go around the city to have more visibility and make the project known.

Figure 4. The van designed for recording the stories.

In the box, one could thus enter it to tell in complete intimacy the aspect of the city they want to share (urban legend, historical fact, a story about an important personality, significant event, etc.). This Box allowed people who have oral narratives around the history of the city, the neighborhood, a character, event, and even legends to tell and share them. All these catches were processed and then disseminated on the day of the event in various places that were visited and open to the public as part of the route organized. Recording interviews was the major step in the project. It made it possible to create all the social and historical content for the programming of the event and the visit route of the monuments.

Figure 5. A person telling their memories in the box (left) and a person telling their memories indoors (right).

 ISBN: 978-1-61208-976-8

B. ICT and media use to promote the heritage

In cultural heritage, people draw upon the digital technologies of social media to represent their heritage independently of museums and archives [11]. In this regard, the activities performed by people can be individual practices, social media presence and the organizational network of NGOs, local traditional media, and public archives constitutes the crowdsourced heritage [12]. In *Ahkili Aliha*, both "conventional" media like Radio and ICT like social media were used to promote the event and engage more people. Social media accounts managed by E&M were used to encourage participation, while radio was used to promote the event.

1) Social media event teaser

A mobilization video was broadcasted via the Edifices et Mémoires Facebook page. The date chosen to lunch the teaser was August 10, 2019, a day before Eid (a religious celebration for Muslims). While the Eid is celebrated across Tunisia, where it's an official day off, it represents great importance for Kairouan due to its Islamic heritage. The event teaser published was accompanied by a text in Arabic and French communicating the concept of the *Ahkili Aliha* project and specifying the dates and the place of recording of the interviews. To reach a maximum number of people, the *Ahkili Aliha* project members chose three influential people to tell the story that links each one of them to the heritage of Kairouan. This aimed to show that heritage belongs to everyone and is not reserved for an elite or experts and to ensure high visibility of the event and encourage people to participate in the recordings. These people appearing in the teaser represent a variety of profiles, including:

- First person: a septuagenarian who lived in the Medina of Kairouan for several years. His testimony presents a "social proof" to encourage the less young and the inhabitants of the Medina to come and tell their stories.
- Second person: a basketball player who is an emblematic figure of Kairouan. His appearance in the video represented a surprise to his fans who expected an interview about a basketball game or his sports career. In the video the player sent a message to the inhabitants of Kairouan to safeguard this heritage that presents their identity. His appearance aroused the interest of young fans of this player in the heritage of Kairouan.
- Third person: a young woman active on social networks. After telling her story, the woman launches a call for participation in the *Ahkili Aliha* project. In the teaser, not putting a recognized expert in the video was decided by the *Ahkili Aliha* project designers (1) to show that heritage is not only limited to experts and (2) to confirm that oral heritage is also cultural heritage that can be shared and promoted thanks to the contributions of the inhabitants.

2) Radio

The concept and objectives of the project, as well as the funding agencies and partners, were presented by the project manager in a morning radio show in August 2019. The purpose of the radio talk for the *Ahkili Aliha* project members was to invite listeners to participate in the week of recording interviews and to announce the date of the closing event. The use of Radio was also to reach an audience who is still using the conventional media channel.

C. Oral heritage promotion for built heritage promotion

Interactive visit routes have been set up, in close collaboration with the partners, highlighting the lived and told stories of the selected neighborhood. Through the visits, 2 main objectives were targeted: First, raise awareness among people of the rich oral heritage being a gateway for a better appropriation and promotion of their built heritage. And second, in the longer term, fight against the destruction of cultural heritage through training in the safeguarding of cultural heritage.

Figure 6. Visitors exploring the old city during the *Ahkili Aliha* event day.

Local musicians performed a concert at the end of the guided tours. Some stalls were set up to sell and promote regional products or local crafts. The aim of the event day

was to create an opportunity to transmit the collective memory from one generation to another.

V. DISCUSSION

A. Non-linearity of Patrimonialization process

The medina of Kairouan is a UNESCO heritage since 1988. To be included on the World Heritage List, sites must be of outstanding universal value and meet at least one out of ten selection criteria. The old city of Kairouan meets five out of the ten selection criteria of UNESCO. We emphasize that the old city of Kairouan has reached at least the step of conservation for two reasons: (1) due to its UNESCO heritage classification and the national and regional efforts to preserve the old city, and (2) the institutionalization, the professionalization, and the practical traductions are confirmed. This leads to four possible scenarios when it comes to the patrimonialization of the old city of Kairouan: either it is in full patrimonialization process, in a de-patrimonialization scenario, or in re-patrimonialization or in an over-patrimonialization scenario.

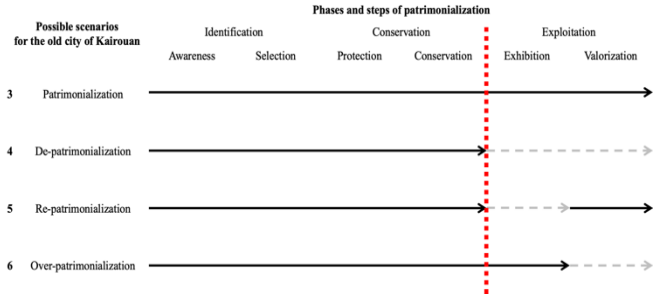

Figure 7. Phases and steps of possible patrimonialization scenarios for the old city of Kairouan.

Based on the participant-observant method, we found out that "celebrating" the city, is a common feeling among its inhabitants on specific religious occasions, such as the month of Ramadan and the celebration of the birthday of the prophet Mohamed. This temporary, but strong connection to the old city is widely present and shared among the participants of *Ahkili Aliha* project. While in the rest of the year, participants said they experienced feelings of belonging to their city, these feelings seem to be not as strong as during the special occasions. This helps us deduct that the patrimonialization process of the old city of Kairouan, is not a constant linear phenomenon. In fact, it seems to be a variant cyclical phenomenon, that depends on different factors, such as the times and the occasions of the connection that the inhabitants feel toward their heritage, and therefore the practices they have in the city and other political events that tend to value, even "glorify" the cities. We cite here the example of the designation of Kairouan as the Islamic Culture Capital for 2009 by The Islamic World Educational, Scientific and Cultural Organization, ICESCO. Once a step is achieved the patrimonialization process is done, it can be also withdrawn or broken if the factor behind it is not present anymore. Hence, we represent the patrimonialization process as a succession of phases, such as Youssef [7]. Every step takes the form of a circle, to represent two scenarios: (1) with the horizontal arrow to signify that the step may lead thereafter but directly to the next step, or (2) with two circular arrows, to signify that the step may be withdrawn and lead to the previous step.

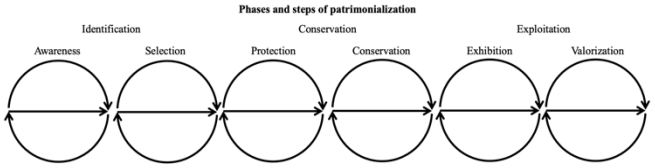

Figure 8. Patrimonialization process in nonlinear representation

While the absence of the steps of the phase of exploitation does not cancel the patrimonialization process, we emphasize that participation and ICT can be a tool to fill the gaps of the missing steps in uncomplete or broken patrimonialization process scenarios.

B. The role of participation and ICT to resume or restore broken exploitation steps in patrimonialization

The new dimensions of cultural practices have generally been articulated around the notions of networking, connectivity, and participation. The new media, such as the ones based on ICT, are a major part of participatory cultural systems [13]. In the patrimonialization process, the actors play different roles in the steps. While the Faro convention puts the people at the heart of the processes of identification, management, and sustainable use of heritage, some steps of the patrimonialization are still exclusive to the actors by action and actors by knowledge (our last work to cite), mainly the protection and conservation steps. While participation guarantees the inclusion of a big number of actors around cultural heritage projects and experiences, it plays an important role in the reconciliation of tangible heritage and intangible heritage. The study case of *Ahkili Aliha* allowed us to show that actors by knowing to contribute more with knowledge related to intangible heritage, such as stories and legends, while the actors by knowledge contribute with their expertise in built heritage, such as information about history and buildings. Participation represents the core concept, with a variety of approaches to follow, in the *Ahkili Aliha* project it is the collaborative action-research. ICT represent the tool that provides a common ground for actors of patrimonialization to meet and share their knowledge. Together, participation and ICT provide an opportunity for cities like Kairouan, to resume and restore their patrimonialization process.

 ISBN: 978-1-61208-976-8

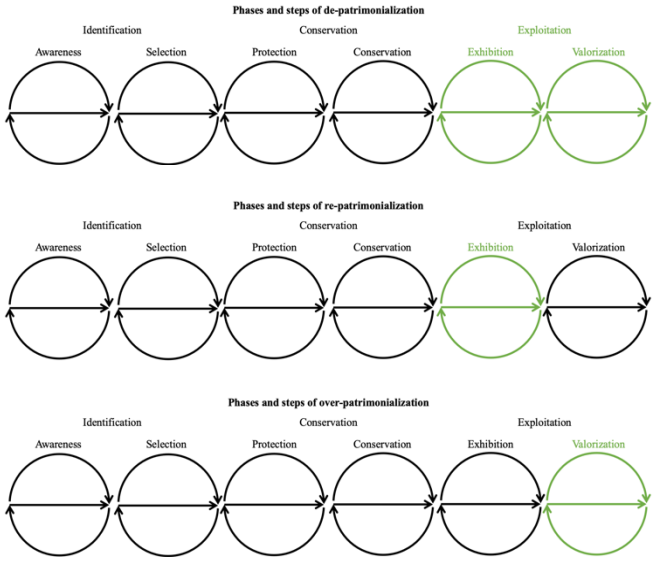

Figure 9. The steps of exhibition and/or valorization (in green) where participation and ICT can play a role to resume uncompleted or suspended or broken patrimonialization process.

VI. CONCLUSION

Contributions. This article has made it possible to highlight the contribution of participation and ICT in a context for the promotion of oral heritage and built heritage in the framework of collaborative action research. It also allowed to underline that the promotion of the built heritage and patrimonial sites cannot be done independently from the promotion of oral heritage and the other constituents of the intangible heritage. The discussion allowed to consider the role of ICT and participation in the steps of exhibition and/or valorization to resume uncompleted or suspended or broken patrimonialization process.

Prospects. A crucial work is to be expected as a result of this project, which consists of analyzing the data collected from the box of *Ahkili Aliha* project. The data captured can represent solid work support for a reflection on patrimonial values of heritage and their links with different actors identified since the heritage value cannot simply be linked to the historical value but its nonquantified values to the individuals and the communities.

When it comes to the future of the *Ahkili Aliha* project, the Regional Commission of Culture had expressed its interest in inviting the project team to present the *Ahkili Aliha* project as part of the Kairouan Intangible Heritage Festival that will take place in April 2020 (unfortunately it was postponed because of the pandemic). The partner associations and citizen volunteers who participated in the *Ahkili Aliha* project also expressed their interest in making the second edition of *Ahkili Aliha* Kairouan in the period of *Mouled*, a religious festivity that the city of Kairouan is known for.

REFERENCES

[1] S. Ben Rajeb and H. Béjar, “Participatory Approach for a "Collaborative Heritage Observatory" in Tunisia”, The Ninth International Conference on Advanced Collaborative Networks, Systems and Applications), IARIA, 2019, pp 12-19, ISBN: 978-1-61208-722-1

[2] J. Eichler “Intangible cultural heritage, inequalities and participation: who decides on heritage?”, The International Journal of Human Rights, 25:5, pp 793-814, Sept. 2020, doi: 10.1080/13642987.2020.1822821

[3] K. Stiti and S. Ben Rajeb, “2W + 1H systematic review to (re)draw Actors and Challenges of Participation(s): Focus on Cultural Heritage”, Architecture, ISSN: 2673-8945

[4] P. Friere, “Pedagogy of the oppressed”, New York: The Continuum International Publishing Group Inc, 1970

[5] F. Kensing and J. Greenbaum, “Heritage: Having a say”, Routledge international handbook of participatory design, pp. 21-36, 2013

[6] E. Amougou, « La question patrimoniale. De la « patrimonialisation » à l’examen des situations concrètes », L’Harmattan, Paris, 282 p, 2004

[7] Z. Youssef and F. Kharrat, « Le processus de patrimonialisation des Médinas de Sousse et Mahdia en Tunisie: vers la reconstitution, l'évaluation et la comparaison », International conférence : Les Médinas à l'époque contemporaine (XX-XXI e siècles): oscillations entre patrimonialisation et marginalisation, Tours, France, 2015

[8] N. K. Denzin, “The research act: A theoretical introduction to sociological methods”, New York: Aldine, p. 291, 1970.

[9] R. K. Yin, “Case study research: Design and methods”, Thousand Oaks, CA: Sage, 1994

[10] M. Minkler and N. Wallerstein, “Community based participatory research for health: Process to outcomes”, 2nd Edition, Jossey Bass, San Francisco, 2008

[11] A. van der Hoeven, “Historic urban landscapes on social media: The contributions of online narrative practices to urban heritage conservation”, City, Culture and Society, Volume 17, pp 61-68, 2019, ISSN: 1877-9166, doi: 10.1016/j.ccs.2018.12.001.

[12] M. Roszczyńska-Kurasińska, A. Domaradzka, B. Ślosarski, and A. Żbikowska, “Facebook Data as Part of Cultural Heritage Investments Toolbox: Pilot Analysis of Users Interests and Preferences Concerning Adaptive Reuse”, Sustainability 13, no. 4: 2410, 202, doi: 10.3390/su13042410

[13] H. Jenkins, “Convergence culture: Where old and new media collide”, New York University Press, 2006